Study of Data Mining and Data Warehousing Challenges With Rdbms

M. B. Bramarambika

Research Scholar, CMJ University, Shillong, Meghalaya

Abstract: ***With the emergence of data warehousing, Decision Support Systems have evolved to its best. At the core of these warehousing systems lies a good database management system. Database server, used for data warehousing, is responsible to provide robust data management, scalability, high performance query processing and integration with other servers. Oracle being the initiator in warehousing servers, provides a wide range of features for facilitating data warehousing. This paper is designed to review the features of data warehousing - conceptualizing the concept of data warehousing and lastly, features of Grade servers for implementing a data warehouse.***

----------♦----------

DATA WAREHOUSE - A CONCEPTUAL OVERVIEW

W.H. Inmon, "father of data warehousing", defined data warehouse as: A data warehouse is a Subject Oriented, Integrated, Non-volatile, and Tine-variant collection of data in support of management's decisions.

With the advancement in the computing technology, the fall in the computer hardware and change in the nature of business - the value of information have raised dramatically. The need of making decisions on the basis of large amount of data, which has the property of diversification along with the hugeness, have raised to a level not comparable to any phase throughout the history of Information Technology. Supplementing was the betterment of server operating systems and the explosion of Internets and Web based applications. The more organized Information database is - the better is the performance of the company. This indispensable requirement to store enormous amount of data lead to the Analytic Systems which in turn gave birth to the idea of Data Warehousing.

Data warehousing is about molding data into information, and storing this information based on the subject rather than application. As mentioned by W.H. Inmon, in one of his artides, the data warehouse environment is the foundation of DSS- Deasion Support Systems.

Going back to the definition of data warehouse, the warehouse is a Subject Oriented, Integrated, Non-volatile, and Time-variant collection of data.

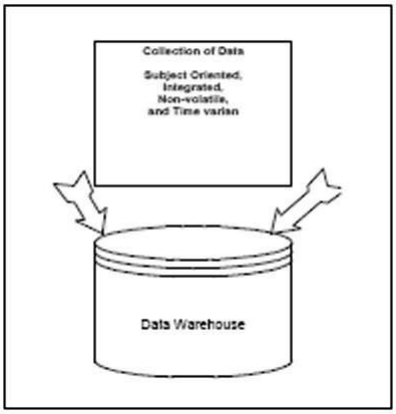


Figure – 1

SUBJECT-ORIENTED

In data warehousing the prime objective of storing data is to facilitate decision process of a company, and within any company data naturally concentrates around subject areas. This leads to the gathering of information around these subjects rather than around the applications or processes.

INTEGRATED

Though the data in the data warehouses is scattered around different tables, databases or even servers but the data is integrated consistently in the values of variables, naming conventions and physical data definitions.

NONVOLATILE

Being the snapshot of operational data on a given specific time, the data in the data warehouses should not be changed or updated - once its loaded from operational system. As the snapshot shows operational data at some moment of time and one expects data warehouse to reflect accurate values of that time frame. There exist only wo operations - the time- based loading of data, accessing the leaded data.

TIME-VARIANT

The value of operational data changes on the basis of time. The time based archival of data from operational systems to data warehouse, makes the value of data, in the data warehouses, being function of time. As data warehouse gives accurate picture of operational data for some gKvn time and the change in the data in warehouse is based on time bised change in operational data, data in the data warehouse is called 'time-variant'.

From the operational systems to the requirement of DSS, to designing of data warehousing, to Implement to ongoing support, data warehousing does not use some alien concepts and is more or less based on the typical System Development Life Cycle (SDLC) concept.

Data warehouses possess a degree of multi-dimensioning in there nature. The advocates of Relational Modeling say that Multi-dimensioning of data is just another way of representation of data in two dimensional relational models. If we agree to the above rationale then the data warehousing comes in the umbrella of traditional RDBMS application development process. Yet indeed, there are some major differences when building a warehouse, including features like hugeness of data or accessibility or providing dynamic access etc. The most important difference is of course the way data is placed in data warehouses, its more like summarized, referenced, de-normalized representation. In short what ever or how ever we develop a data warehouse it should at least be capable of providing ad hoc complex, statistical, and analytical queries to facilitate decision making process.

ARCHITECTURE OF DATA WAREHOUSE

As repeatedly mentioned in this paper, the prime concern of providing a separate set of data - the data warehouse, is to facilitate Business .Analysts in the process of Decision Making. Essentially data warehousing is the "warehousing" data outside operational systems and this has not significantly changed with the evolution of data warehousing systems. Prime reason of this separation is that the evaluation and analysis, done by analysts, require complex and analytic queries - the effect of which is the performance degradation of operational systems. Another important feature is the combination of data from more than one operational system to provide the ability of cross-referencing.

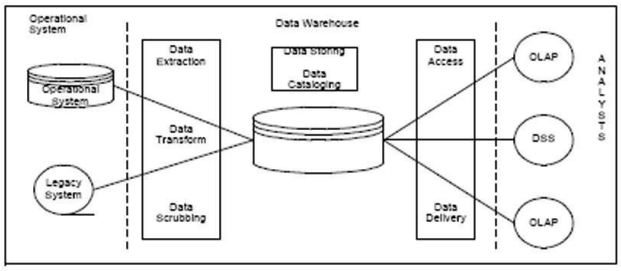


Figure - 2

Most of the data warehousing done, passes three-tier architecture.

The base level from which data is extracted is operational system (OLTP) and the legacy systems, from which data is transformed and loaded into the warehouse database. So the middle level is the data warehouse and the top most level is the analytic system (OLAP) and Decision Support System (DSS). OLAP systems utilize the data warehouse to provide multi-dimensional view. Functionally a data warehouse can he divided into following:

i Data Extraction

ii Transformation and Scrubbing

iii Storing and Cataloging

iv Data Access

v Data Delivery

All of the above functions are self-explanatory. The process starts with the extraction of data from operational system and legacy systems, then comes the transformation and cleaning of data during this process summarization and aggregation is also done. Data Storage represents the process of storing transformed and cleaned data in a relational database. Data access holds the query processing multi-dimensional analysis and data mining. Lastly comes the function of data delivery to the end-users, which may be the part of data warehouse or can come under the umbrella of OLAPs.

A data warehouse, being unique in the class of applications, possesses a structure, which is different from other database applications. Being used for analytic purpose it is designed in a way so that it can facilitate complex queries. Mostly the business analysts focus on the summarized data, time variant data. So the data warehouses are designed to facilitate the above process. Data warehouses hold different levels of summarization

and details. It also has wo groups of detail data, the current detail data and older detail data.

Current detail data, reflecting the most current happening in the organization, is highly voluminous and is always stored on disk storage. It may reach space as much as gigabytes or even terabytes. The reason of Iving so sizable is that it asserts lowest level of granularity.

Old detail data, as the name shows, is the data, which is not that frequently, used. Due to the infrequent requirement this data is stored on some cheaper storage mediums like tape cartridge.

Then comes the level of summarization, the difference between lightly summarized and highly summarized is quite obvious. Lightly summarized data is the summary of detailed granulized data. Whereas highly summarized data is more compact than summarized and is based on lightly summarized figures. Both of these reside on disk media, as these are accessed very frequendy.

Meta data, being very important data repository, resides on different dimension than other data classes. As it may he accessed by any of the other layers and work as a linkage warehouse and operational environment.

WAREHOUSE DATABASE SERVER - ITS ROLE IN DATA WAREHOUSING

Data in the data warehouse database is organized by subject rather than applications or processes, and this data is extracted and refreshed from operational system on a periodic basis. We have already discussed the three-tiered architecture in which first tier is the operational system, middle is the data warehouse database server and last one is front-ended client applications, including DSS and OLAP applications. In three-tiered architecture the warehouse database server works as the heart of warehouse application. Though a simpler form of data warehouse applications exist, in which the architecture is two tiered - tierl includes the Operational System as well as Warehouse database and tier2 is the client front-end Decision Support applications.

One can't disagree to the fact that Database servers are at the core of every application that supports business decisions, sped ally data warehouses - providing robust data management and scalable, high-performance query processing.

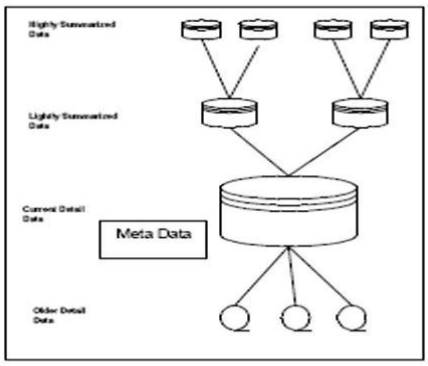


Figure – 3

warehouse servers are categorized in two types, RDBMS (Relational Database) and MDD (Mulu-dimensional Database) - the choice is based on type of data stored in warehouse.

RDBMS is based on the concept of mathematical relation operation. The implementation of RDBMS is based on two- dimensional relationship of related data - called the tables. Whereas, MDD can be viewed as cube, where information is pilled on various axes of cube. Taking as an example, the case of Sales production of a company - Sales are related to salespersons, the geographical region, and some ume frame, this result in three-dimensional view of data. The cross- section of these three can give the required data. However MDD just work with finite set of data and information which is highly related to each other.

Relational database technology has an edge on MDD, when we are considering huge data storage capadty or portability issue or security. RDBMS is an old and proven technology in data storage and recovery. MDD is popular for its Instance Response, Implementation ease, and integration with Meta-data. Either we choose MDD or RDBMS in both cases a database server has a very central role in the data warehouse architecture.

DATA MODELING - STAR SCHEMA AS CHOICE

Data modeling, the process of miking data models is not unique for warehousing; in fact we use this tool in the development of all kinds of database applications. The reason in data warehousing is pretty much the same.

1. Defining the scope of data warehouse

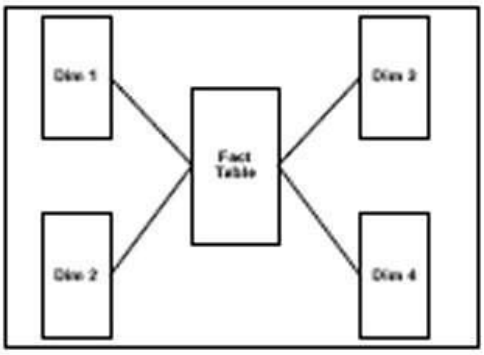
2. Viewing the complexity of the relationship between data

3. Recognizing and controlling redundancy

Decision-makers during the analysis generally formulate complex queries, which are based on multiple dimensions. As data warehousing is done to facilitate this type of multi-dimensional queries for decision making, the modeling of data also tends to bear multiple dimensions. Data warehousing done using relational database technology generally holds modeling in star schema.

Star schema is the implementation of multiple dimensions in the relational modeling. This schema addresses data navigation difficulty and its dimensions are the categories bv which analysts organize the information. Star schema at the lowest level is the relationship between tables, but with the expansion of scope of warehouse the model tends to become more and more complex So it's a good practice to aggregate data into levels of hierarchy. The relationship among different objects is provided by introducing fact tables - tables having primary key compounded by primary keys of all dimensions.

The fact table is the central table in star architecture containing the data links or points to establish dimensions in different entities. Technically, this table is just intersection of entity primary keys.



META DATA - THE DATA ABOUT DATA

Meta data provides data repository. Providing both technical and business view of data stored in the data warehouse. It lavs out the physical structures that includes:

1. Data elements and their types
2. Business definition for the data elements
3. How to update data and on which frequency
4. Different data elements having same meaning
5. Valid values for each data elements

Meta data plays very important rule in the definition, building, management and maintenance of data warehouses. In a data warehouse Meta data are categorized into Business and Technical Meta data. Business Meta data describes what's in the warehouse, its meaning - in business terms. The business Meta data lies above technical Meta data, adding some more details to the extracted material. This type of Meta data is important as it facilitates business users and increases the accessibility. In contrast, technical Meta data describes the data elements as they exist in the warehouse. This type of Meta data is used for data modeling, initially, and once the warehouse is erected this Meta data is frequendy used by warehouse administrator and software tools.

USING THE INTERNET TECHNOLOGY

With the technology provided through Internets, the transfer of information has become very easy. On the other hand is the requirement of accessing data warehouse globally. Putting both together gives a very effective solution to give the access to data warehouses on the global scale. To provide global access to a data warehouse using web is like giving easy access to data on the whenever, whoever basis. However, there are some issues which has to be sorted before the effective utilization of internet technology, like the usage of web server along with the database server, security issues, and some issues like providing ways for query and report purpose.

ORACLE - A CHOICE FOR IMPLEMENTING DATA WAREHOUSE

Since the conceptualization of data warehouse, many database venders have tried to mold their database systems for accommodating it. Amongst which was Oracle that systematically evolved to address spedfic needs of warehousing. When considering a data warehouse implemented in a RDBMS, there are some technology requirements like query processing, data storage, scalability, integration with other systems and lastly the security management.

QUERY PROCESSING

Queries in a data warehouse generally involve very large amount of data. Also it's not rare to find complex operations like multi-table joins, sorting and aggregation in data warehouse queries. These operations are generally set-oriented; operating on some groups of records based on specified criteria. Most of the queries in dedsion making process are multi-dimensional in nature, based on star schema. .Another important feature in query processing of data warehouse is that queries are not pre-defined and are based on the business-users runtime

criteria.

Features like query optimization, access and joining methods and parallel execution of queries are very vital for performance of data warehouse.

DATA MANAGEMENT AND SCALABILITY

This is the way data is leaded, organized, stored, accessed and maintained in a database. The database operations such as data loading enforcing constraints, building indexes, collecting statistics on the data, reorganizing tables and indexes, building aggregates or summaries, and data purging are included in data management. Its not unusual to find wrv large databases when implementing a data warehouse, also the growth of a warehouse is in big data leaps. The database operations, listed above, are functions of database size.

To effectively meet the needs of data warehousing the database server has to provide capacity to deal with large data volumes and data operations should also be tuned for the same reason.

While the scope of data warehouse is not at all limited, this feature leads to the scalability of both users and data. With the globalization of organizations - number of end-users, requiring to use warehouse, have increased dramatically. The supporting of this population of users is the responsibility of database server. This include supporting wide range of hardware, operating systems, and clients - trying to access data warehouse from widely apart physical positioning.

When considering scale of data, server has to support data volumes of gigabytes, terabytes or even beyond. The scalability doesn't merely mean the capacity to store immense data; it encompasses the ability to efficiendy process queries, the capability to perform data management operations, and delivering business-critical availability, all at huge scale.

INTEGRATION WITH ODIER SYSTEMS

In the process of decision making, the analysts have to access data even beyond the boundaries of operational data and its not always wise to transfer each bit of data from systems like this to data warehouse. So database servers should provide provisions to link the warehouse application to systems - like SAP, BAAN or PeopleSoft

SECURITY MANAGEMENT

With the physical size of data warehouses and number of users requiring to access data warehouse in the process of decision miking - the security of organization's critical data is at stake if database server is not able to manage security properly.

ORACLE SERVER - WHERE DOES IT STAND

Oracle has been amongst the earlier database management systems extending its features to accommodate data warehouse related features. It was the era of Oracle? when the concept of data warehouse came and Oracle-corporation right away recognized its importance. Oracle v7.3 provided features like parallel query execution, parallel data management, cost-based query optimization, efficient bitmap indexing and hash joining emtedded in query execution. Then came Orade8 - enhancing the features, already provided by Oracle \'7.3- Linking the server with tools like Oracle Discoverer and Oracle express haw made Oracle the must viable option for data warehousing. Below we are going to discuss features provided in Orade to enhance the server capabilities for the implementation of data warehouses.

QUERY PROCESSING

Oracle? advanced its architecture to improve the Query Optimizer as well as the execution of query.

QUERY OPTIMIZATION: The main task of query optimizer is to choose the most efficient way to execute a SQL statement - the DML (Data Manipulation Language) are considered for optimization. Oracle produces an execution plan for the optimization purpose. Oracle Optimizer takes following steps for the selection of best execution plan:

Evaluation of query expression and modification as per required. The optimizer assesses expressions construct and whenever required introduces some modification to enhance the speed and reduce resource utilization. Some examples of which are given in Figure 5-

```
{Where Clause}
  cl_name like 'xyz'
cl_name = 'xyz'
  cl_name in ('a', 'b', 'c')
cl_name='a' or cl_name='b' or
cl_name='c'
  cl_name > any (select amount
                 from payment
                 where place = 'xxx')
exists (select amount
        from payment
        where place = 'xxx'
        and cl_name > amount)
  cl_name > all (select amount
                 from payment
```

Figure-5 (Bold shows the optimize query)

Transformation of complex and symbiotic queries into equivalent joins statements. In the process of transformation, optimizer modifies two types of queries; queries containing OR to UNION .ALL and complex queries into join statements.

For queries having views - the optimizer merges the query statement with that of view. Examples of such optimization is given in Figure 6

```
create or replace view view1
as
select c1_1, c1_2
from   table1
where  c11 > 10;

(when selecting from view1)
select c12
from   view1
where  c12 > 15;

(optimizer modifies the query into)
select c12
from   table1
where  c11 > 10
and    c12 > 15;
```

Figure – 6

Selection of Optimization approach from Rule-Based Optimization and Cost-Based Optimization. Rule-based approach chooses execution path based on heuristicallv ranked operations. When more than one execution paths exists, rule-based approach selects path with lower rank. Cost-based appioach optimizes a query based on following steps:

1. Firstly, all potential execution plans are predetermined by optimizer - plans are based on access paths.

2. Then, optimizer estimates the ccst of each execution plan based on the data distribution and storage characteristic statistics - the statistics are based on table structure, indexes and clusters, I/O and CPU rime, the available memory.

3. Lastly, optimizer compares the cost of execution plans and selects one with lowest cost.

The selections of appropriate access path when a query is based on more than one table. Generally there exists more than one access paths when the table-data is accessed. The optimizer chooses the most appropriate access path based on the Rule-based or Cost-based approach.

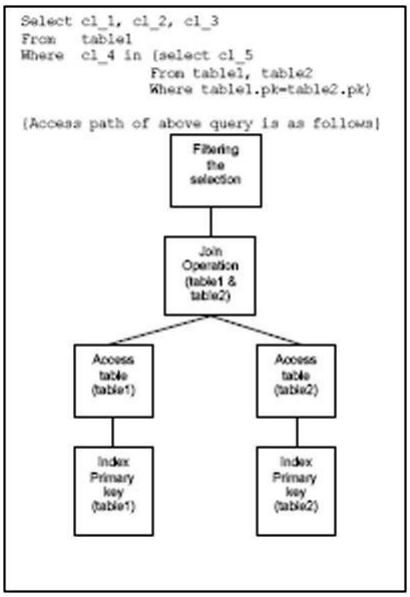


Figure – 7

When joining more than wo schemas, optimizer decides which pair to join first.

QUERY EXECUTION: Introducing the intra-query parallelism via Parallel Query option. Oracle? provided parallel execution of complex queries having SQL operations like; SELECT, sub-queries in INSERT. DELETE OR UPDATE. CREATE TABLE based on sub-query, and CREATE INDEX commands. The parallel Query option improves the performance of data manipulation operations in very large databases, like warehouses. Best performance can be viewed on SMP (Symmetric Multiprocessor) and MPP (Massively Parallel Processing) machines. The query writers have to implicitly command the parallel query option and also declare degree of parallelism. Figure 8 explain how parallel query-works in Oracle.

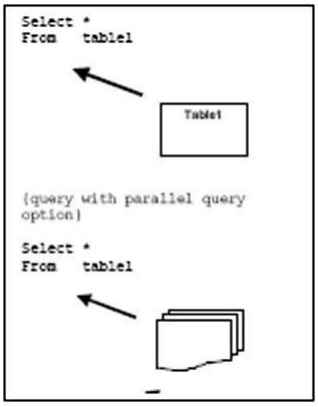


Figure – 8 (Parallelism of degree 3)

In order to provide parallelism in query execution many initial parameters have to be configured. Once the system is configured to run queries with parallel query option, it is the task of Query Coordiniror Process to initiate parallel query- servers and coordinate between the results from these query servers. The number of query servers, running in parallel to complete one operation, is called Degree of Parallelism.

ORACLE WAREHOUSE ARCHITECTURE

This architecture is designed using the RDBMS server and tools provided by Oracle. The Oracle warehouse can be developed using two-tiered or three-tiered architecture. The two-tiered architecture involves the database server at back- end and front-end decision support tools. A more complex warehouse involves separate tiers tor data access from operational source, data storage and presentation of data for decision support.

Tier 1 - Accessing Source Data

Data can be accessed from multiple sources*, including operational systems, Legacy systems and other Oracle applications. Utilities like SQL'Loader, export/import Oracle schemas, SQL Stored Procedures can be used tor the data transfer. For transferring data from legacy systems and a wide range ot other systems Oracle Transparent Gateways are used.

Tier 2 - The Server for Warehouse

1 he server for data warehousing can be of RDBMS or MDD type. Oracle provides solution for both options. If one decides to use Multi-Dimensional Database Architecture then there exist Oracle Express Server, and for Relation Database Architecture the option is Oracle7, 0racle8 or 0racle8i. Warehouse can be designed by integrating the two.

Tier 3 - Decision Support System

To entertain Business Analysts, both DSS and OLAP tools can be provided in the data warehouse. With tools like Oracle Reports, Oracle Discoverer and Oracle Express, users can have access on data warehouse on the whenever and however basis.

REFERENCES

- W.H.Inmon - known as father of Data Warehousing What is Data Warehouse?
- Vivek R. Gupta - Senior Consultant, Services corporation, Chicago, Illinois. An Introduction to Data Warehousing David Heise - CIO Andrews University Data Warehousing at Avondale College
- Dr. James Goodnight - CEO SAS Institute Inc.
- Data Warehousing: Understanding Its Role in a Business Management .Architecture
- Oracle Publications
- Oracle? - Server Concept Manual
- Oracle Publications
- Oracle7 - Server Tuning Manual
- Oracle Publications Oracled Concepts
- Oracle White Paper, June 1997, Oracled for Data Warehousing
- Oracle White Paper, June 1997,
- Oracled - Enabling Decisions in the New Business Era
- Oracle White Paper, June 1997,
- Oracled - The Database for Network Computing
- Oracle White Paper, June 1997, Star Queries in Oracled
- Oracle White Paper, June 1997, Oracled - The Hub of Oracle Warehouse
- Winter Corporation White Paper
- Large Scale Data Warehousing with OracleSi
- Oracle White Paper, November 1998, Oracledi - The Database for Internet Computing